Electronic Version
Stylesheet Version v1.1.1

Description

BIOINFORMATICALLY DETECTABLE GROUP OF NOVEL REGULATORY GENES AND USES THEREOF

BACKGROUND OF INVENTION

CONTINUATION STATEMENT

[0001] This application is a continuation of U.S Provisional Patent Application Serial No. 60468251, filed 07-May-03, entitled "Bioinformatically Detectable Group of Novel Regulatory Genes and Uses Thereof ", and is a continuation of U.S Patent Application Serial No.10604985, filed 29-Aug-03, entitled "Bioinformatically Detectable Group of Novel Regulatory Genes and Uses Thereof", and is a continuation in part of U.S Provisional Patent Application Serial No. 10345201, filed 16-Jan-03, entitled "Bioinformatically Detectable Group of Novel Regulatory Genes and Uses Thereof", and is a continuation in part of U.S Patent Application Serial No. 10293338, filed 14-Nov-03, entitled "Bioinformatically Detectable Group of Novel Regulatory Genes and Uses Thereof ", and is a continuation in part of U.S Patent Application Serial No. 10310914, filed 06-Dec-02, entitled "Bioinformatically Detectable Group of Novel Regulatory Genes and Uses Thereof ", and is a continuation in part of U.S Patent Application Serial No. 10321503, filed 18-Dec-02, entitled "Bioinformatically Detectable Group of Novel Regulatory Genes and Uses Thereof ", and is a continuation in part of U.S Patent Application filed 29-Aug-03, entitled "Bioinformatically Detectable Group of Novel Regulatory Genes and Uses of Thereof", and is a continuation in part of U.S Patent Application Serial No. 10604926 , filed 27-Aug-03, entitled "Bioinformatically Detectable Group of Novel Regulatory Genes and Uses Thereof ", and is a continuation in part of U.S Patent Application , filed 28-Aug-03, entitled "Bioinformatically Detectable Group of Novel Regulatory Genes and Uses Thereof ", and is a continuation in part of U.S Patent Application Serial No.10604726, filed 13-Aug-03, entitled "Bioinformatically Detectable Group of Novel Regulatory Genes and Uses Thereof", and is a continuation in part of U.S Patent Application Serial No.10604727, filed 13-Aug-03, entitled "Bioinformatically Detectable Group of Novel Regulatory Genes and Uses Thereof", the disclosures of which applications are all hereby incorporated by reference and claims priority therefrom.

FIELD OF THE INVENTION

[0002] The present invention relates to a group of bioinformatically detectable novel genes, here identified as "genomic address messenger" or "GAM" genes, which are believed to be related to the micro RNA (miRNA) group of genes.

DESCRIPTION OF PRIOR ART

[0003] Small RNAs are known to perform diverse cellular functions, including post-transcriptional gene expression regulation. The first two such RNA genes, Lin-4 and Let-7, were identified by genetic analysis of Caenorhabditis Elegans (Elegans) developmental timing, and were termed short temporal RNA (stRNA) (Wightman, B., Ha, I., Ruvkun, G., Cell 75, 855 (1993); Erdmann, V.A.. et al., Nucleic Acids Res. 29, 189 (2001); Lee, R. C., Feinbaum, R. L., Ambros, V., Cell 75, 843 (1993); Reinhart, B. et al., Nature 403, 901 (2000)).

[0004] Lin-4 and Let-7 each transcribe a ~22 nucleotide (nt) RNA, which acts a post transcriptional repressor of target mRNAs, by binding to elements in the 3"-untranslated region (UTR) of these target mRNAs, which are complimentary to the 22 nt sequence of Lin-4 and Let-7 respectively. While Lin-4 and Let-7 are expressed at different developmental stage, first larval stage and fourth larval stage respectively, both specify the temporal progression of cell fates, by triggering post-transcriptional control over other genes (Wightman, B., Ha, I., Ruvkun, G., Cell 75, 855 (1993); Slack et al., Mol.Cell 5 ,659 (2000)). Let-7 as well as its temporal regulation have been demonstrated to be conserved in all major groups of bilaterally symmetrical animals, from nematodes, through flies to humans (Pasquinelli, A., et al. Nature 408 ,86 (2000)).

[0005] The initial transcription product of Lin-4 and Let-7 is a ~60-80nt RNA, the nucleotide sequence of the first half of which is partially complimentary to that of its second half, therefore allowing this RNA to fold onto itself, forming a "hairpin structure". The final gene product is a ~22nt RNA, which is "diced" from the above mentioned "hairpin structure", by an enzyme called Dicer, which also apparently also mediates the complimentary binding of this ~22nt segment to a binding site in the 3" UTR of its target gene.

[0006] Recent studies have uncovered 93 new genes in this class, now referred to as micro RNA or miRNA genes, in genomes of Elegans, Drosophilea, and Human (Lagos-Quintana, M., Rauhut, R., Lendeckel, W., Tuschl, T., Science 294 ,853 (2001); Lau, N.C., Lim, L.P., Weinstein, E.G., Bartel, D.P., Science 294 ,858 (2001); Lee, R.C., Ambros, V., Science 294 ,862 (2001). Like the well studied Lin-4 and Let-7, all newly found MIR genes produce a ~60-80nt RNA having a nucleotide sequence capable of forming a "hairpin structure". Expressions of the precursor ~60-80nt RNA and of the resulting diced ~22nt RNA of most of these newly discovered MIR genes have been detected.

[0007] Based on the striking homology of the newly discovered MIR genes to their well-studied predecessors Lin-4 and Let-7, the new MIR genes are believed to have a similar basic function as that of Lin-4 and Let-7: modulation of target genes by complimentary binding to the UTR of these target genes, with special emphasis on modulation of developmental control processes. This is despite the fact that the above mentioned recent studies did not find target genes to which the newly discovered MIR genes complementarily bind. While existing evidence suggests that the number of regulatory RNA genes "may turn out to be very large, numbering in the hundreds or even thousands in each genome", detecting such genes is challenging (Ruvkun G., "Perspective: Glimpses of a tiny RNA world", Science 294 ,779 (2001)).

[0008] The ability to detect novel RNA genes is limited by the methodologies used to detect such genes. All RNA genes identified so far either present a visibly discernable whole body phenotype, as do Lin-4 and Let-7 (Wightman et. al., Cell 75, 855 (1993); Reinhart et al., Nature 403, 901 (2000)), or produce significant enough quantities of RNA so as to be detected by the standard biochemical genomic techniques, as do the 93 recently detected miRNA genes. Since a limited number clones were sequenced by the researchers discovering these genes, 300 by Bartel and 100 by Tuschl (Bartel et. al., Science 294 ,858 (2001); Tuschl et. al., Science 294 ,853 (2001)), the RNA genes found can not be much rarer than 1% of all RNA genes. The recently detected miRNA genes therefore represent the more prevalent among the miRNA gene family.

[0009] Current methodology has therefore been unable to detect RNA genes which either do not present a visually discernable whole body phenotype, or are rare (e.g. rarer than 0.1% of all RNA genes), and therefore do not produce significant enough quantities of RNA so as to be detected by standard biochemical technique.

BRIEF DESCRIPTION OF SEQUENCE LISTING, LARGE TABLES AND COMPUTER PROGRAM LISTING

[0010] Sequence listing attached to the present invention. The Sequence listing comprising 664,408 genomic sequences, and is contained in a file named SEQ_LIST.TXT (101,307KB).

[0011] Large tables relating to genomic sequences are stored in 8 files, each comprising a respective one of the following table files: TABLE1.TXT (219,587KB); TABLE2.TXT (1,805KB); TABLE3.TXT (10,543KB); TABLE4.TXT (2,023KB), TABLE5.TXT (391,401KB), TABLE6.TXT (1,197,508KB); TABLE7.TXT (1,078,187KB) and TABLE8 (43,997KB).TXT.

SUMMARY OF INVENTION

[0012] The present invention relates to a novel group of regulatory, non-protein coding genes, which are functional in specifically inhibiting translation of other genes, some of which are known to be involved in various diseases. Each gene in this novel group of genes, here identified as "GAM" or "Genomic Address Messengers", specifically inhibits translation of one of more other "target" genes by means of complimentary hybridization of a segment of the RNA transcript encoded by GAM, to an inhibitor site located in an untranslated region (UTR) of the mRNA of the one or more "target" genes.

[0013] In various preferred embodiments, the present invention seeks to provide improved method and system for specific modulation of expression of specific known "target" genes involved in significant human diseases, and improved method and system for detection of expression of these target genes.

[0014] Accordingly, the invention provides several substantially pure DNAs (e.g., genomic DNA, cDNA or synthetic DNA) each encoding a novel gene of the GAM group of gene, vectors comprising the DNAs, probes comprising the DNAs, a method and system for selectively modulating translation of known "target" genes utilizing the vectors, and a method and system for detecting expression of known "target" genes utilizing the probe.

[0015] By "substantially pure DNA" is meant DNA that is free of the genes which, in the naturally-occurring genome of the organism from which the DNA of the invention is derived, flank the genes discovered and isolated by the present invention. The term therefore includes, for example, a recombinant DNA which is incorporated into a vector, into an autonomously replicating plasmid or virus, or into the genomic DNA of a prokaryote or eukaryote at a site other than its natural site; or which exists as a separate molecule (e.g., a cDNA or a genomic or cDNA fragment produced by PCR or restriction endonuclease digestion) independent of other sequences. It also includes a recombinant DNA which is part of a hybrid gene encoding additional polypeptide sequence.

[0016] "Inhibiting translation" is defined as the ability to prevent synthesis of a specific protein encoded by a respective gene, by means of inhibiting the translation of the mRNA of this gene. "Translation inhibiter site" is defined as the minimal DNA sequence sufficient to inhibit translation.

[0017] There is thus provided in accordance with a preferred embodiment of the present invention a bioinformatically detectable novel gene encoding substantially pure DNA wherein: RNA encoded by the bioinformatically detectable novel gene is about 18 to about 24 nucleotides in length, and originates from an RNA precursor, which RNA precursor is about 50 to about 120 nucleotides in length, a nucleotide sequence of a first half of the RNA precursor is a partial inversed-reversed sequence of a nucleotide sequence of a second half thereof, a nucleotide sequence of the RNA encoded by the novel gene is a partial inversed-reversed sequence of a nucleotide sequence of a binding site associated with at least one target gene, the novel gene cannot be detected by either of the following: a visually discernable whole body phenotype, and detection of 99.9% of RNA species shorter than 25 nucleotides expressed in a tissue sample, and a function of the novel gene is bioinformatically deducible.

[0018] There is further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable novel gene encoding substantially pure DNA wherein: RNA encoded by the bioninformatically detectable novel gene includes a plurality of RNA sections, each of the RNA sections being about 50 to about 120 nucleotides in length, and including an RNA segment, which RNA segment is about 18 to about 24 nucleotides in length, a nucleotide sequence of a first half of each of the RNA sections encoded by the novel gene is a partial inversed-reversed sequence of nucleotide sequence of a second half thereof, a nucleotide sequence of each of the RNA segments encoded by the novel gene is a partial inversed-reversed sequence of the nucleotide sequence of a binding site associated with at least one target gene, and a function of the novel gene is bioinformatically deducible from the following data elements: the nucleotide sequence of the RNA encoded by the novel gene, a nucleotide sequence of the at least one target gene, and function of the at least one target gene.

[0019] There is still further provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable novel gene encoding substantially pure DNA wherein: RNA encoded by the bioinformatically detectable novel gene is about 18 to about 24 nucleotides in length, and originates from an RNA precursor, which RNA precursor is about 50 to about 120 nucleotides in length, a nucleotide sequence of a first half of the RNA precursor is a partial inversed-reversed sequence of a nucleotide sequence of a second half thereof, a nucleotide sequence of the RNA encoded by the novel gene is a partial inversed-reversed sequence of a nucleotide sequence of a binding site associated with at least one target gene, a function of the novel gene is modulation of expression of the at least one target gene, and the at least one target gene does not encode a protein.

[0020] There is additionaly provided in accordance with another preferred embodiment of the present invention A bioinformatically detectable novel gene encoding substantially pure DNA wherein: the bioinformatically detectable novel gene does not encode a protein, RNA encoded by the bioinformatically detectable novel gene is maternally transferred by a cell to at least one daughter cell of the cell, a function of the novel gene includes modulation of a cell type of the daughter cell, and the modulation is bioinformatically deducible.

[0021] There is moreover provided in accordance with another preferred embodiment of the present invention a bioinformatically detectable novel gene encoding substantially pure DNA wherein: the bioinformatically detectable novel gene does not encode a protein, a function of the novel gene is promotion of expression of the at lease one target gene, and the at least one target gene is bioinformatically deducible.

[0022] Further in accordance with a preferred embodiment of the present invention the function of the novel gene is bioinformatically deducible from the following data elements: the nucleotide sequence of the RNA encoded by the bioinformatically detectable novel gene, a nucleotide sequence of the at least one target gene, and a function of the at least one target gene.

[0023] Still further in accordance with a preferred embodiment of the present invention the RNA encoded by the novel gene complementarily binds the binding site associated with the at least one target gene, thereby modulating expression of the at least one target gene.

[0024] Additionally in accordance with a preferred embodiment of the present invention the binding site associated with at least one target gene is located in an untranslated region of RNA encoded by the at least one target gene.

[0025] Moreover in accordance with a preferred embodiment of the present invention the function of the novel gene is selective inhibition of translation of the at least one target gene, which selective inhibition includes complementary hybridization of the RNA encoded by the novel gene to the binding site.

[0026] Further in accordance with a preferred embodiment of the present invention the invention includes a vector including the DNA.

[0027] Still further in accordance with a preferred embodiment of the present invention the invention includes a method of selectively inhibiting translation of at least one gene, including introducing the vector.

[0028] Moreover in accordance with a preferred embodiment of the present invention the introducing includes utilizing RNAi pathway.

[0029] Additionally in accordance with a preferred embodiment of the present invention the invention includes a gene expression inhibition system including: the vector, and a vector inserter, functional to insert the vector into a cell, thereby selectively inhibiting translation of at least one gene.

[0030] Further in accordance with a preferred embodiment of the present invention the invention includes a probe including the DNA.

[0031] Still further in accordance with a preferred embodiment of the present invention the invention includes a method of selectively detecting expression of at least one gene, including using the probe.

[0032] Additionally in accordance with a preferred embodiment of the present invention the invention includes a gene expression detection system including: the probe, and a gene expression detector functional to selectively detect expression of at least one gene.

BRIEF DESCRIPTION OF DRAWINGS

[0033] Fig.1 is a simplified diagram illustrating the genomic differentiation enigma that the present invention addresses;

[0034] Figs. 2 through 4 are schematic diagrams which when taken together provide an analogy that illustrates a conceptual model of the present invention, addressing the genomic differentiation enigma;

[0035] Figs. 5A and 5B are schematic diagrams, which when taken together illustrate a "genomic records" concept of the conceptual model of the present invention, addressing the genomic differentiation enigma;

[0036] Fig. 6 is a schematic diagram illustrating a "genomically programmed cell differentiation" concept of the conceptual model of the present invention, addressing the genomic differentiation enigma;

[0037] Fig. 7 is a schematic diagram illustrating a "genomically programmed cell-specific protein expression modulation" concept of the conceptual model of the present invention, addressing the genomic differentiation enigma;

[0038] Fig. 8 is a simplified diagram describing a mode by which genes of a novel group of genes of the present invention, modulates expression of known target genes;

[0039] Fig. 9 is a simplified block diagram illustrating a bioinformatic gene detection system capable of detecting genes of the novel group of genes of the present invention, which system is constructed and operative in accordance with a preferred embodiment of the present invention;

[0040] Fig. 10 is a simplified flowchart illustrating operation of a mechanism for training of a computer system to recognize the novel genes of the present invention, which mechanism is constructed and operative in accordance with a preferred embodiment of the present invention;

[0041] Fig. 11A is a simplified block diagram of a non-coding genomic sequence detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0042] Fig. 11B is a simplified flowchart illustrating operation of a non-coding genomic sequence detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0043] Fig. 12A is a simplified block diagram of a hairpin detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0044] Fig. 12B is a simplified flowchart illustrating operation of a hairpin detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0045] Fig. 13A is a simplified block diagram of a dicer-cut location detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0046] Fig. 13B is a simplified flowchart illustrating training of a dicer-cut location detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0047] Fig. 13C is a simplified flowchart illustrating operation of a dicer-cut location detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0048] Fig. 14A is a simplified block diagram of a target-gene binding-site detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0049] Fig. 14B is a simplified flowchart illustrating operation of a target-gene binding-site detector constructed and operative in accordance with a preferred embodiment of the present invention;

[0050] Fig. 15 is a simplified flowchart illustrating operation of a function & utility analyzer constructed and operative in accordance with a preferred embodiment of the present invention;

[0051] Fig. 16 is a simplified diagram describing a novel bioinformatically detected group of regulatory genes, referred to here as Genomic Record (GR) genes, each of which encodes an □operon-like□ cluster of novel micro RNA-like genes, which in turn modulates expression of plurality of target genes;

[0052] Fig. 17 is a simplified diagram illustrating a mode by which genes of a novel group of operon-like genes of the present invention, modulate expression of other such genes, in a cascading manner;

[0053] Fig. 18 is a block diagram illustrating an overview of a methodology for finding novel genes and novel operon-like genes of the present invention, and their respective functions;

[0054] Fig. 19 is a block diagram illustrating different utilities of novel genes and novel operon-like genes, both of the present invention;

[0055] Figs. 20A and 20B are simplified diagrams, which when taken together illustrate a mode of gene therapy applicable to novel genes of the present invention;

[0056] Fig. 21A is an annotated sequence of EST72223 comprising known miRNA gene MIR98 and novel gene GAM25, both detected by the gene detection system of the present invention;

[0057] Figs. 21B and 21C are pictures of laboratory results demonstrating laboratory confirmation of expression of known gene MIR98 and of novel bioinformatically detected gene GAM25 respectively, both of Fig. 21A, thus validating the bioinformatic gene detection system of the present invention;

[0058] Fig. 21D provides pictures of laboratory results, which when taken together demonstrate further laboratory confirmation of expression of the bioinformatically detected novel gene GAM25 of Fig. 21A, and another novel bioinformatically detected gene GAM26, thus further validating the bioinformatic gene detection system of the present invention;

[0059] Fig. 22A is an annotated sequence of an EST7929020 comprising novel genes GAM24 and GAM26 detected by the gene detection system of the present invention;

[0060] Fig. 22B is a picture of laboratory results, which confirm expression of bioinformatically detected novel genes GAM24 and GAM26 of Fig. 22A, thus further validating the bioinformatic gene detection system of the present invention;

[0061] Fig. 22C is a picture of laboratory results, which confirm endogenous expression of bioinformatically detected novel gene GAM26 of Fig. 22A;

[0062] Fig. 23A is an annotated sequence of an EST1388749 comprising novel gene GAM27 detected by the gene detection system of the present invention;

[0063] Fig. 23B is a picture of laboratory results, which confirm expression of the bioinformatically detected novel gene GAM27 of Fig. 23A;

Brief Description of Sequences

[0064] A Sequence Listing of genomic sequences of the present invention designated SEQ ID:1 through SEQ ID:664008 is attached to this application. The genomic listing comprises the following nucleotide sequences: Genomic sequences designated SEQ ID:1 through SEQ ID:23013 are nucleotide sequences of 23,013 gene precursors of respective novel genes of the present invention; Genomic sequences designated SEQ ID:23014 through SEQ ID:50760 are nucleotide sequences of 27,747 genes of the present invention; and Genomic sequences designated SEQ ID:50761 through SEQ ID:664008 are nucleotide sequences of 613,248 target gene binding sites.

Detailed Description

[0065] Reference is now made to Fig. 1 which is a simplified diagram providing a conceptual explanation of a genomic differentiation enigma, which the present invention addresses.

[0066] Fig. 1 depicts different cell types in an organism, such as CARTILAGE CELL, LIVER CELL, FIBROBLAST CELL and BONE CELL all containing identical DNA, and deriving from the initial FERTILIZED EGG CELL, and yet each of these cells expressing different proteins, and hence acquiring different shape and function.

[0067] The present invention proposes that the inevitable conclusion from this constraint is, however, strikingly simple: the coding system used must be modular. It must comprise multiple modules, or records, one for each cell-type, and a mechanism whereby each cell at its inception is instructed which record to open, and behaves according to instructions in that record.

[0068] This modular code concept is somewhat difficult to grasp, since we are strongly habituated to viewing things from an external viewpoint. An architect, for example, looks at a blueprint of a building, which details exactly where each element (block, window, door, electrical switch, etc.) is to be placed relative to all other elements, and then instructs builders to place these elements in their designated places. This is an external viewpoint: the architect is external to the blueprint, which itself is external to the physical building, and its different elements. The architect may therefore act as an "external organizing agent": seeing the full picture and the relationships between all elements, and being able to instruct from the outside where to place each of them.

[0069] Genomics differentiation coding evidently works differently, without any such external organizing agent: It comprises only one smart block (the first cell), which is the architect and the blueprint, and which continuously duplicates itself, somehow knowing when to manifest itself as a block and when as a window, door, or electrical switch.

[0070] Reference is now made to Figs. 2 through 4 which are schematic diagrams which when taken together provide an analogy that illustrates a conceptual model of the present invention, addressing the genomic differentiation enigma.

[0071] Reference is now made to Fig. 2A. Imagine a very talented chef, capable of preparing any meal provided he is given specific written cooking instructions. This chef is equipped with two items: (a) a thick recipe book, and (b) a small note with a number scribbled on it. The book comprises multiple pages, each page detailing how to prepare a specific meal. The small note indicates the page to be opened, and therefore the meal to be prepared. The chef looks at the page-number written on the note, opens the recipe book at the appropriate page, and prepares the meal according to the written instructions on this page. As an example, Fig. 2A depicts a CHEF holding a note with the number 12 written on it, he opens the book on page 12, and since that page contains the recipe for preparing BREAD, the CHEF prepares a loaf of BREAD.

[0072] Reference is now made to Fig. 2B, which depicts two identical chefs, CHEF A and CHEF B, holding an identical recipe book. Despite their identity, and the identity of their recipe book, since CHEF A holds a note numbered 12, and therefore opens the book on page 12 and prepares BREAD, whereas CHEF B holds a note numbered 34 and therefore opens the book on page 34 and prepares a PIE.

[0073] Reference is now made to Fig. 3. Imagine the chef of the analogy is also capable of duplicating himself once he has finished preparing the specified meal. The format of the book is such that at the bottom of each page, two numbers are written. When he has finished preparing the meal specified on that page, the chef is trained to do the following: (i) divide himself into two identical duplicate chefs, (ii) duplicate the recipe book and hand a copy to each of his duplicate chefs, and (iii) write down the two numbers found at the bottom of the page of the meal he prepared, on two small notes, handing one note to each of his two duplicate chefs.

[0074] Each of the two resulting duplicate chefs are now equipped with the same book, and have the same talent to prepare any meal, but since each of them received a different note, they will now prepare different meals.

[0075] Fig. 3 depicts CHEF A holding a recipe book and receiving a note numbered 12. CHEF A therefore opens the book on page 12 and prepares BREAD. When he is finished making bread, CHEF A performs the following actions: (i) divides himself into two duplicate chefs, designated CHEF B and CHEF C, (ii) duplicates his recipe book handing a copy to each of CHEF B and CHEF C, (iii) writes down the numbers found at the bottom of page 12, numbers 34 and 57, on two notes, handing note numbered 34 to CHEF B and note numbered 57 to CHEF C.

[0076] Accordingly, CHEF B receives a note numbered 34 and therefore opens the recipe book on page 34 and prepares PIE, whereas CHEF C receives a note numbered 57 and therefore opens the book on page 57 and therefore prepares RICE.

[0077] It is appreciated that while CHEF A, CHEF B & CHEF C are identical and hold identical recipe books, they each prepare a different meal. It is also appreciated that the meals prepared by CHEF B and CHEF C are determined CHEF A, and are mediated by the differently numbered notes passed on from CHEF A to CHEF B and CHEF C.

[0078] It is further appreciated that the mechanism illustrated by Fig. 3 enables an unlimited lineage of chefs to divide into duplicate, identical chefs and to determine the meals those duplicate chefs would prepare. For example, having been directed to page 34, when CHEF B divides into duplicate chefs (not shown), he will instruct its two duplicate chefs to prepare meals specified on pages 14 and 93 respectively, according to the numbers at the bottom of page 34 to which he was directed. Similarly, CHEF C will instruct its duplicate chefs to prepare meals specified on pages 21 and 46 respectively, etc.

[0079] Reference is now made to Fig. 4. Imagine that the cooking instructions on each page of the recipe book are written in shorthand format: The main meal-page to which the chef was directed by the scribbled note, merely contains a list of numbers which direct him to multiple successive pages, each specifying how to prepare an ingredient of that meal.

[0080] As an example, Fig. 4 depicts CHEF A of FIGS 2 and 3, holding a recipe book and a note numbered 12. Accordingly, CHEF A opens the recipe book on page 12, which details the instructions for preparing BREAD. However, the "instructions" on making BREAD found on page 12 comprise only of 3 numbers, 18, 7 and 83, which "refer" CHEF A to pages detailing preparation of the ingredients of BREAD FLOUR, MILK and SALT, respectively.

[0081] As illustrated in Fig. 4, turning from the main "meal page" (e.g. 12) to respective "ingredients pages" (e.g. pages 18, 7 & 83) is mediated by scribbled notes with the page-numbers written on them. In this analogy, the scribbled notes are required for seeking the target pages to be turned to both when turning to main "meal pages" (e.g. page 12), as well as when turning to "ingredient pages" (e.g. pages 18, 7 & 83).

[0082] The chef in the given analogy, schematically depicted in FIGS 2 through 4, represents a cell; the thick recipe book represents the DNA; preparing a meal in the given analogy represents the cell manifesting itself as a specific cell-type; and ingredients of a meal represent proteins expressed by that cell-type. Like the chef equipped with the thick recipe book in the given analogy, all cells in an organism contain the same DNA and are therefore each potentially capable of manifesting itself as any cell-type, expressing proteins typical of that cell type.

[0083] Reference is now made to Figs. 5A and 5B which are schematic diagrams, which when taken together illustrate a "genomic records" concept of the conceptual model of the present invention, addressing the genomic differentiation enigma.

[0084] The Genomic Records concept asserts that the DNA (the thick recipe book in the illustration) comprises a very large number of Genomic Records (analogous to pages in the recipe book), each containing the instructions for differentiation of a different cell-type, or developmental process. Each Genomic Record is headed by a very short genomic sequence which functions as a "Genomic Address" of that Genomic Record (analogous to the page number in the recipe book). At its inception, in addition to the DNA, each cell also receives a short RNA segment (the scribbled note in the illustration). This short RNA segment binds complementarily to a "Genomic Address" sequence of one of the Genomic Records, thereby activating that Genomic Record, and accordingly determining the cell"s-fate (analogous to opening the book on the page corresponding to the number on the scribbled note, thereby determining the meal to be prepared).

[0085] Reference is now made to Fig. 5A. Fig 5A illustrates a CELL which comprises a GENOME. The GENOME comprises a plurality of GENOMIC RECORDS, each of which correlates to a specific cell type (for clarity only 6 sample genomic records are shown). Each genomic record comprises genomic instructions on differentiation into a specific cell-type, as further elaborated below with reference to Fig. 7. At cell inception, the CELL receives a maternal short RNA segment, which activates one of the GENOMIC RECORDS, causing the cell to differentiate according to the instructions comprised in that genomic record. As an example, Fig. 5A illustrates reception of a maternal short RNA segment designated A" and outlined by a broken line, which activates the FIBRO genomic record, causing the cell to differentiate into a FIBROBLAST CELL.

[0086] Reference is now made to Fig. 5B, which is a simplified schematic diagram, illustrating cellular differentiation mediated by the "Genomic Records" concept. Fig. 5B depicts 2 cells in an organism, designated CELL A and CELL B, each having a GENOME. It is appreciated that since CELL A and CELL B are cells in the same organism, the GENOME of

CELL A is identical to that of CELL B. Despite having an identical GENOME, CELL A differentiates differently from CELL B, due to activation of different genomic records in these two cells. In CELL A the FIBRO GENOMIC RECORD is activated, causing CELL A to differentiate into a FIBROBLAST CELL, whereas in CELL B the BONE GENOMIC RECORD is activated, causing the CELL B to differentiate into a BONE CELL. The cause for activation of different genomic records in these two cells is the different maternal short RNA which they both received: CELL A received a maternal short RNA segment designated A" which activated genomic record FIBRO, whereas CELL B received a maternal short RNA segment designated B" which activated genomic record BONE.

[0087] Reference is now made to Fig. 6 which is a schematic diagram illustrating a "genomically programmed cell differentiation" concept of the conceptual model of the present invention, addressing the genomic differentiation enigma.

[0088] A cell designated CELL A divides into 2 cells designated CELL B and CELL C. CELL A, CELL B and CELL C each comprise a GENOME, which GENOME comprises a plurality of GENOMIC RECORDS. It is appreciated that since CELL A, CELL B and CELL C are cells in the same organism, the

GENOME of these cells, and the GENOMIC RECORDS comprised therein, are identical.

[0089] As described above with reference to Fig. 5B, at its inception, CELL A receives a maternal short RNA segment, designated A" and marked by a broken line, which activates the FIBRO genomic record, thereby causing CELL A to differentiate into a FIBROBLAST CELL. However, Fig. 6 shows further details of the genomic records: each cell genomic record also comprises two short genomic sequences, referred to here as Daughter Cell Genomic Addresses. Blocks designated B and C are Daughter Cell Genomic Addresses of the FIBRO Genomic Record. At cell division, each parent cell transcribes two short RNA segments, corresponding to the two Daughter Cell Genomic Addresses of the Genomic Record of that parent cell, and transfers one to each of its two daughter cells. CELL A of Fig. 6 transcribes and transfers to its two respective daughter cells, two short RNA segments, outlined by a broken line and designated B" and C", corresponding to daughter cell genomic addresses designated B and C comprised in the FIBRO genomic record.

[0090] CELL B therefore receives the above mentioned maternal short RNA segment designated B", which binds complementarily to genomic address designated B of genomic record BONE, thereby activating this genomic record, which in turn causes CELL B to differentiate into a BONE CELL. Similarly, CELL C receives the above mentioned maternal short RNA segment designated C", which binds complementarily to genomic address designated C of genomic record CARTIL., thereby activating this genomic record, which in turn causes CELL C to differentiate into a CARTILAGE CELL.

[0091] It is appreciated that the mechanism illustrated by Fig. 6 enables an unlimited lineage of cells to divide into daughter cells containing the same DNA, and to determine the cell-fate of these daughter cells. For example, when CELL B and CELL C divide into their respective daughter cells (not shown), they will transfer short RNA segments designated D" & E", and F" & G" respectively, to their respective daughter cells. The cell fate of each of these daughter cells would be determined by the identity of the maternal short RNA segment they receive, which would determine the genomic record activated.

[0092] Reference is now made to Fig. 7 which is a schematic diagram illustrating a "genomically programmed cell-specific protein expression modulation" concept of the conceptual model of the present invention, addressing the genomic differentiation enigma.

[0093] Cell A receives a maternal short RNA segment designated A", which activates a genomic record designated FIBRO, by anti-sense binding to a binding site "header" of this genomic record, designated A. Genomic record FIBRO encodes 3 short RNA segments, designated 1, 2 and 4 respectively, which modulate expression of target genes designated GENE1, GENE2 and GENE4 respectively. Modulation of expression of these genes results in CELL A differentiating into a FIBROBLAST CELL.

[0094] Reference is now made to Fig. 8 which is a simplified diagram illustrating a mode by which genes of a novel group of genes of the present invention, modulate expression of known target genes.

[0095] The novel genes of the present invention are micro RNA (miRNA)-like, regulatory RNA genes, modulating expression of known target genes. This mode of modulation is common to other known miRNA genes, as described hereinabove with reference to the background of the invention section.

[0096] GAM GENE and GAM TARGET GENE are two human genes contained in the DNA of the human genome.

[0097] GAM GENE encodes a GAM PRECURSOR RNA. However, similar to other miRNA genes, and unlike most ordinary genes, its RNA, GAM PRECURSOR RNA, does not encode a protein.

[0098] GAM PRECURSOR RNA folds onto itself, forming GAM FOLDED PRECURSOR RNA. As Fig.8 illustrates, GAM FOLDED PRECURSOR RNA forms a "hairpin structure", folding onto itself. As is well known in the art, this "hairpin structure", is typical genes of the miRNA genes, and is due to the fact that nucleotide sequence of the first half of the RNA of a gene in this group is an accurate or partial inversed-reversed sequence of the nucleotide sequence of its second half. By "inversed-reversed" is meant a sequence which is reversed and wherein each nucleotide is replaced by a complimentary nucleotide, as is well known in the art (e.g. ATGGC is the inversed-reversed sequence of GCCAT).

[0099] An enzyme complex, designated DICER COMPLEX, "dices" the GAM FOLDED PRECURSOR RNA into a single stranded RNA segment, about 22 nucleotides long, designated GAM RNA. As is known in the art, "dicing" of the hairpin structured RNA precursor into shorter RNA segments about 22 nucleotides long by a Dicer type enzyme is catalyzed by an enzyme complex comprising an enzyme called Dicer together with other necessary proteins.

[0100] GAM TARGET GENE encodes a corresponding messenger RNA, designated GAM TARGET RNA. This GAM TARGET RNA comprises 3 regions: a 5" untranslated region, a protein coding region and a 3" untranslated region, designated 5"UTR, PROTEIN CODING and 3"UTR respectively.

[0101] GAM RNA binds complementarily a BINDING SITE, located on the 3"UTR segment of GAM TARGET RNA. This complementarily binding is due to the fact that the nucleotide sequence of GAM RNA is an accurate or partial inversed-reversed sequence of the nucleotide sequence of BINDING SITE.

[0102] The complementary binding of GAM RNA to BINDING SITE inhibits translation of GAM TARGET RNA into GAM TARGET PROTEIN. GAM TARGET PROTEIN is therefore outlined by a broken line.

[0103] It is appreciated by one skilled in the art that the mode of translational inhibition illustrated by Fig. 8 with specific reference to GAM genes of the present invention is in fact common to all other miRNA genes. A specific complimentary binding site has been demonstrated only for Lin-4 and Let-7. All the other newly discovered miRNA genes (over 300) are also believed by those skilled in the art to modulate expression of other genes by complimentary binding, although specific complimentary binding sites for these genes have not yet been found (Ruvkun G., "Perspective: Glimpses of a tiny RNA world", Science 294 ,779 (2001)). The present invention discloses a novel group of genes, the GAM genes, belonging to the miRNA genes group, and for which a specific complimentary binding has been determined.

[0104] Table 1, hereby incorporated by reference, provides detailed descriptions of each of a plurality of GAM GENEs as described generally by Fig.8.

[0105] Nucleotide sequences of each of a plurality of GAM GENEs described by Fig. 8 and their respective genomic source and chromosomal location are further described hereinbelow with reference to Table 2, hereby incorporated by reference.

[0106] Nucleotide sequences of GAM PRECURSOR RNA, and a schematic representation of a predicted secondary folding of GAM FOLDED PRECURSOR RNA, of each of a plurality of GAM GENEs described by Fig. 8 are further described hereinbelow with reference to Table 3, hereby incorporated by reference.

[0107] Nucleotide sequences of a `diced` GAM RNA of each of a plurality of GAM GENEs described by Fig. 8 are further described hereinbelow with reference to Table 4, hereby incorporated by reference.

[0108] Nucleotide sequences of target binding sites, such as BINDING SITE-I, BINDING SITE-II and BINDING SITE-III of Fig. 8, found on GAM TARGET RNA, of each of a plurality of GAM GENEs described by Fig. 8, and schematic representation of the complementarity of each of these target binding sites to each of a plurality of GAM RNA described by Fig. 8 are described hereinbelow with reference to Table 5, hereby incorporated by reference.

[0109] It is appreciated that specific functions and accordingly utilities of each of a plurality of GAM GENEs described by Fig. 8 correlate with, and may be deduced from, the identity of the GAM TARGET GENEs that each of said plurality of GAM GENEs binds and inhibits, and the function of each of said GAM TARGET GENEs, as elaborated hereinbelow with reference to Table 6, hereby incorporated by reference.

[0110] Studies establishing known functions of each of a plurality of TARGET GENEs of GAM GENEs of Fig. 8, and correlation of said each of a plurality of TARGET GENEs to known diseases are listed in Table 7, and are hereby incorporated by reference.

[0111] The present invention discloses a novel group of genes, the GAM genes, belonging to the miRNA genes group, and for which a specific complementary binding has been determined.

[0112] Reference is now made to Fig. 9 which is a simplified block diagram illustrating a bioinformatic gene detection system capable of detecting genes of the novel group of genes of the present invention, which system is constructed and operative in accordance with a preferred embodiment of the present invention.

[0113] A centerpiece of the present invention is a bioinformatic gene detection engine 100, which is a preferred implementation of a mechanism capable of bioinformatically detecting genes of the novel group of genes of the present invention.

[0114] The function of the bioinformatic gene detection engine 100 is as follows: it receives three types of input, expressed RNA data 102, sequenced DNA data 104, and protein function data 106, performs a complex process of analysis of this data as elaborated below, and based on this analysis produces output of a bioinformatically detected group of novel genes designated 108.

[0115] Expressed RNA data 102 comprises published expressed sequence tags (EST) data, published mRNA data, as well as other sources of published RNA data. Sequenced DNA data 104 comprises alphanumeric data describing sequenced genomic data, which preferably includes annotation data such as location of known protein coding regions relative to the sequenced data. Protein function data 106 comprises scientific publications reporting studies which elucidated physiological function known proteins, and their connection, involvement and possible utility in treatment and diagnosis of various diseases. Expressed RNA data 102, sequenced DNA data 104 may preferably be obtained from data published by the National Center for Bioinformatics (NCBI) at the National Institute of Health (NIH), as well as from various other published data sources. Protein function data 106 may preferably be obtained from any one of numerous relevant published data sources, such as the Online Mendelian Inherited Disease In Man (OMIM(TM)) database developed by John Hopkins University, and also published by NCBI.

[0116] Prior to actual detection of bioinformatically detected novel genes 108 by the bioinformatic gene detection engine 100, a process of bioinformatic gene detection engine training & validation designated 110 takes place. This process uses the known miRNA genes as a training set (some 200 such genes have been found to date using biological laboratory means), to train the bioinformatic gene detection engine 100 to bioinformatically recognize miRNA-like genes, and their respective potential target binding sites. Bioinformatic gene detection engine training & validation 110 is further describe hereinbelow with reference to Fig. 10.

[0117] The bioinformatic gene detection engine 100 comprises several modules which are preferably activated sequentially, and are described as follows:

[0118] A non-coding genomic sequence detector 112 operative to bioinformatically detect non-protein coding genomic sequences. The non-coding genomic sequence detector 112 is further described hereinbelow with reference to Figs. 11A and 11B.

[0119] A hairpin detector 114 operative to bioinformatically detect genomic "hairpin-shaped" sequences, similar to GAM FOLDED PRECURSOR of Fig. 8. The hairpin detector 114 is further described hereinbelow with reference to Figs. 12A and 12B.

[0120] A dicer-cut location detector 116 operative to bioinformatically detect the location on a hairpin shaped sequence which is enzymatically cut by DICER COMPLEX of Fig. 8. The dicer-cut location detector 116 is further described hereinbelow with reference to Fig. 13A.

[0121] A target-gene binding-site detector 118 operative to bioinformatically detect target genes having binding sites, the nucleotide sequence of which is partially complementary to that of a given genomic sequence, such as a sequence cut by DICER COMPLEX of Fig. 8. The target-gene binding-site detector 118 is further described hereinbelow with reference to Figs. 14A and 14B.

[0122] A function & utility analyzer 120 operative to analyze function and utility of target genes, in order to identify target genes which have a significant clinical function and utility. The function & utility analyzer 120 is further described hereinbelow with reference to Fig. 15.

[0123] Hardware implementation of the bioinformatic gene detection engine 100 is important, since significant computing power is preferably required in order to perform the computation of bioinformatic gene detection engine 100 in reasonable time and cost.

[0124] For example, it is estimated that a using a powerful

8-processor server (e.g. DELL POWEREDGE (TM) 8450, 8 XEON (TM) 550MHz processors, 8 GB RAM), over 6 years (!) of computing time are required to detect all MIR genes in the human EST data, together with their respective binding sites.

[0125] Various computer hardware and software configurations may be utilized in order to address this computation challenge, as is known in the art. A preferred embodiment of the present invention may preferably comprise a hardware configuration, comprising a cluster of one hundred PCs (PENTIUM (TM) IV, 1.7GHz, with 40GB storage each), connected by Ethernet to 7 servers (2-CPU, XEON (TM) 1.2-2.2GHz, with ~200GB storage each), combined with an 8-processor server (8-CPU, Xeon 550Mhz w/ 8GB RAM) connected via 2 HBA fiber-channels to an EMC CLARIION (TM) 100-disks, 3.6 Terabyte storage device. A preferred embodiment of the present invention may also preferably comprise a software configuration which utilizes a commercial database software program, such as MICROSOFT (TM) SQL Server 2000. Using such preferred hardware and software configuration, may reduce computing time required to detect all MIR genes in the human EST data, and their respective binding sites, from 6 years to 45 days.

[0126] It is appreciated that the abovementioned hardware configuration is not meant to be limiting, and is given as an illustration only. The present invention may be implemented in a wide variety of hardware and software configurations.

[0127] The present invention discloses 23,053 novel genes of the GAM group of genes, which have been detected bioinformatically, as described hereinbelow , and 7,399 novel genes of the GR group of genes, which have been detected bioinformatically. Laboratory confirmation of 4 bioinformatically predicted genes of the GAM group of genes, and 2 bioinformatically predicted genes of the GR group of genes, is described hereinbelow with reference to Figs. 21 through 23.

[0128] Reference is now made to Fig. 10 which is a simplified flowchart illustrating operation of a mechanism for training of a computer system to recognize the novel genes of the present invention. This mechanism is a preferred implementation of the bioinformatic gene detection engine training & validation 110 described hereinabove with reference to Fig. 9.

[0129] Bioinformatic gene detection engine training & validation 110 of Fig. 9 begins by training the bioinformatic gene detection engine to recognize known miRNA genes, as designated by numeral 122. This training step comprises hairpin detector training & validation 124, further described hereinbelow with reference to Fig. 12 A, dicer-cut location detector training & validation 126, further described hereinbelow with reference to Fig. 13A and 13B, and target-gene binding-site detector training & validation 128, further described hereinbelow with reference to Fig. 14A.

[0130] Next, the bioinformatic gene detection engine 100 is used to bioinformatically detect sample novel genes, as designated by numeral 130. An example of a sample novel gene thus detected is described hereinbelow with reference to Fig. 21.

[0131] Finally, wet lab experiments are preferably conducted in order to validate expression and preferably function the sample novel genes detected by the bioinformatic gene detection engine 100 in the previous step. An example of wet-lab validation of the abovementioned sample novel gene bioinformatically detected by the system is described hereinbelow with reference to Figs. 22A and 22B.

[0132] Reference is now made to Fig. 11A which is a simplified block diagram of a preferred implementation of the non-coding genomic sequence detector 112 described hereinabove with reference to Fig. 9. Non-protein coding genomic sequence detector 112 of Fig. 9 preferably receives as input at least two types of published genomic data: expressed RNA data 102, including EST data and mRNA data, and sequenced DNA data 104. After its initial training, indicated by numeral 134, and based on the above-mentioned input data, the non-protein coding genomic sequence detector 112 produces as output a plurality of non-protein coding genomic sequences 136. Preferred operation of the non-protein coding genomic sequence detector 112 is described hereinbelow with reference to Fig. 11B.

[0133] Reference is now made to Fig. 11B which is a simplified flowchart illustrating a preferred operation of the non-coding genomic sequence detector 112 of Fig. 9. Detection of non-protein coding genomic sequences to be further analyzed by the system generally preferably progresses in one of the following two paths.

[0134] A first path for detecting non-protein coding genomic sequences begins by receiving a plurality of known RNA sequences, such as EST data. Each RNA sequence is first compared to all known protein-coding sequences, in order to select only those RNA sequences which are non-protein coding. This can preferably be performed by BLAST comparison of the RNA sequence to known protein coding sequences. The abovementioned BLAST comparison to the DNA preferably also provides the localization of the RNA on the DNA.

[0135] Optionally, an attempt may be made to "expand" the non-protein RNA sequences thus found, by searching for transcription start and end signals, upstream and downstream of location of the RNA on the DNA respectively, as is well known in the art.

[0136] A second path for detecting non-protein coding genomic sequences starts by receiving DNA sequences. The DNA sequences are parsed into non protein coding sequences, based on published DNA annotation data: extracting those DNA sequences which are between known protein coding sequences. Next, transcription start and end signals are sought. If such signals are found, and depending on their "strength", probable expressed non-protein coding genomic sequences are yielded.

[0137] Reference is now made to Fig. 12A which is a simplified block diagram of a preferred implementation of the hairpin detector 114 described hereinabove with reference to

Fig. 9.

[0138] The goal of the hairpin detector 114 is to detect "hairpin" shaped genomic sequences, similar to those of known miRNA genes. As mentioned hereinabove with reference to Fig. 8, a "hairpin" genomic sequence refers to a genomic sequence which "folds onto itself" forming a hairpin like shape, due to the fact that nucleotide sequence of the first half of the nucleotide sequence is an accurate or

[0139] The hairpin detector 114 of Fig. 9 receives as input a plurality of non-protein coding genomic sequences 136 of Fig. 11A, and after a phase of hairpin detector training & validation 124 of Fig. 10, is operative to detect and output "hairpin shaped" sequences found in the input expressed non-protein coding sequences, designated by numeral 138.

[0140] The phase of hairpin detector training & validation 124 is an iterative process of applying the hairpin detector 114 to known hairpin shaped miRNA genes, calibrating the hairpin detector 114 such that it identifies the training set of known hairpins, as well as sequences which are similar thereto. Preferred operation of the hairpin detector 114 is described hereinbelow with reference to Fig. 12B.

[0141] Reference is now made to Fig. 12B which is a simplified flowchart illustrating a preferred operation of the hairpin detector 114 of Fig. 9.

[0142] A hairpin structure is a two dimensional folding structure, resulting from the nucleotide sequence pattern: the nucleotide sequence of the first half of the hairpin sequence is an inversed-reversed sequence of the second half thereof. Different methodologies are known in the art for detection of various two dimensional and three dimensional hairpin structures.

[0143] In a preferred embodiment of the present invention, the hairpin detector 114 initially calculates possible 2-dimensional (2D) folding patterns of a given one of the non-protein coding genomic sequences 136, preferably using a 2D folding algorithm based on free-energy calculation, such as the Zucker algorithm, as is well known in the art.

[0144] Next, the hairpin detector 114 analyzes the results of the 2D folding, in order to determine the presence, and location of hairpin structures. A 2D folding algorithm typically provides as output a listing of the base-pairing of the 2D folded shape, i.e. a listing of which all two pairs of nucleotides in the sequence which will bond. The goal of this second step, is to asses this base-pairing listing, in order to determine if it describes a hairpin type bonding pattern.

[0145] The hairpin detector 114 then assess those hairpin structures found by the previous step, comparing them to hairpins of known miRNA genes, using various parameters such as length, free-energy, amount and type of mismatches, etc. Only hairpins that bear statistically significant resemblance of the population of hairpins of known miRNAs, according to the abovementioned parameters are accepted.

[0146] Lastly, the hairpin detector 114 attempts to select those hairpin structures which are as stable as the hairpins of know miRNA genes. This may be achieved in various manners. A preferred embodiment of the present invention utilizes the following methodology comprising three steps:

[0147] First, the hairpin detector 114 attempts to group potential hairpins into "families" of closely related hairpins. As is known in the art, a free-energy calculation algorithm, typically provides multiple "versions" each describing a different possible 2D folding pattern for the given genomic sequence, and the free energy of such possible folding. The hairpin detector 114 therefore preferably assesses all hairpins found on all "versions", grouping hairpins which appear in different versions, but which share near identical locations into a common "family" of hairpins. For example, all hairpins in different versions, the center of which is within 7 nucleotides of each other may preferably be grouped to a single "family".

[0148] Next, hairpin "families" are assessed, in order to select only those families which represent hairpins that are as stable as those of known miRNA hairpins. For example, preferably only families which are represented in at least 65% of the free-energy calculation 2D folding versions, are considered stable.

[0149] Finally, an attempt is made to select the most suitable hairpin from each selected family. For example, preferably the hairpin which appears in more versions than other hairpins, and in versions the free-energy of which is lower, may be selected.

[0150] Reference is now made to Fig. 13A which is a simplified block diagram of a preferred implementation of the dicer-cut location detector 116 described hereinabove with reference to Fig. 9.

[0151] The goal of the dicer-cut location detector 116 is to detect the location in which DICER COMPLEX of Fig. 8, comprising the enzyme Dicer, would "dice" the given hairpin sequence, similar to GAM FOLDED PRECURSOR RNA, yielding GAM RNA both of Fig. 8.

[0152] The dicer-cut location detector 116 of Fig. 9 therefore receives as input a plurality of hairpins on genomic sequences 138 of Fig. 12A, which were calculated by the previous step, and after a phase of dicer-cut location detector training & validation 126 of Fig. 10, is operative to detect a respective plurality of dicer-cut sequences from hairpins 140, one for each hairpin.

[0153] In a preferred embodiment of the present invention, the dicer-cut location detector 116 preferably uses a Support Vector Machine (SVM) trained on the known dicer-cut locations of known miRNA genes, in order to predict dicer-cut locations of novel GAM genes. Dicer-cut location detector training & validation 126, which is further described hereinbelow with reference to Fig. 13B.

[0154] Reference is now made to Fig. 13B which is a simplified flowchart illustrating a preferred implementation of dicer-cut location detector training & validation 126 of Fig. 13A.

[0155] The general goal of the dicer-cut location detector training & validation 126 is to analyze known hairpin shaped miRNA-precursors and their respective dicer-cut miRNA, in order to determine a common pattern to the dicer-cut location of the known miRNA genes. Once such a common pattern is deduced, it may preferably be used by the dicer cut location detector 116, in detecting the predicted dicer-cut sequences from hairpins 140, from the respective hairpins on genomic sequences 138, all of Fig. 13A.

[0156] First, the dicer-cut location of all known miRNA genes is obtained and studied, so as to train the dicer cut location detector 116: for each of the known miRNA, the location of the miRNA relative to its hairpin-shaped miRNA-precursor is noted.

[0157] The 3" and 5" ends of the dicer-cut location for each of the known miRNA genes is represented relative to , which is known for known miRNA genes, is noted relative to the above, as well as to the nucleotides in each location along the hairpin. Frequency of identity of nucleotides, and nucleotide-pairing, relative to their location in the hairpin, and relative to the known dicer-cut location in the known miRNA genes is analyzed and modeled.

[0158] Different techniques are well known in the art for analysis of existing pattern from a given "training set" of species belonging to a genus, which techniques are then capable, to a certain degree, to detect similar patterns in other species not belonging to the training-set genus. Such techniques include, but are not limited to neural networks, Bayesian networks, Support Vector Machines (SVM), Genetic Algorithms, Markovian modeling, and others, as is well known in the art.

[0159] The dicer-cut location detector module uses standard machine learning techniques to predict the 5" and 3" ends of the miRNA excised, or "diced" by the Dicer enzyme from the miRNA hairpin shaped precursor, based on known pairs of miRNA-precursors and their respective resulting miRNAs. The nucleotide sequences of 128 distinct published miRNA were used for training and evaluation of the dicer-cut location detector module. These included 112 MIR genes for which sequences of both final miRNA and miRNA precursor sequences were published, and 16 MIR genes, for which only the final miRNA sequence was published, without the miRNA precursor sequence. In case of the latter, approximations of the precursor hairpin sequences were manually assessed from the DNA sequence in which the MIR was located, using the Zucker-Turner algorithm. The 128 known miRNAs were split into two disjoint sets, used for training and for evaluation respectively.

[0160] Using the abovementioned training set, a Support Vector Machine (SVM) predictor was implemented, which tests every possible nucleotide on a hairpin as a candidate for being the 5" end or the 3" end of a miRNA. The training set of the published miRNA precursor sequences was used for training two separate SVM classifiers, one which produces a model for the 5" end of a miRNA relative to its hairpin precursor, and one which produces a similar model for its 3" end. The models take into account the distance of the respective (3" or 5") end of the miRNA from the hairpin"s loop, the nucleotides at its vicinity and the local "bulge" (i.e. base-pair mismatch) structure.

[0161] Performance of the resulting predictor, evaluated on the abovementioned validation set of 64 published miRNAs, was found to be as follows: in 70% of known miRNAs 5"-end location was determined within up to 2 nucleotides.

[0162] Reference is now made to Fig. 13C which is a simplified flowchart illustrating operation of dicer-cut location detector 116 of Fig. 9, constructed and operative in accordance with a preferred embodiment of the present invention.

[0163] The dicer cut location detector 116 is a machine learning computer program module, which is trained on recognizing dicer-cut location of known miRNA genes, and based on this training, is operable to detect dicer cut location of novel GAM FOLDED PRECURSOR RNA. In a preferred embodiment of the present invention, the dicer-cut location module preferably utilizes a Support Vector Machine (SVM) approach, as is well known in the art.

[0164] When assessing a novel GAM precursor, all 19-24 nucleotide long segments comprised in the GAM precursor are initially considered as "potential GAMs", since the dicer-cut location is initially unknown.

[0165] For each such potential GAM, its 3" end is scored by the 3" end recognition classifier, and its 5" end is scored by the 5" end recognition classifier.

[0166] Scores of the abovementioned two classifiers are integrated, yielding an integrated score for each "potential GAM".

[0167] The integrated score is then evaluated as follows: (a) the "potential GAM" having the highest score is taken to be the most probably GAM, and (b) if the integrated score of this "potential GAM" is higher than a pre-defined threshold, then the potential GAM is accepted as the PREDICTED GAM.

[0168] Reference is now made to Fig. 14A which is a simplified block diagram of a preferred implementation of the target-gene binding-site detector 118 described hereinabove with reference to Fig. 9. The goal of the target-gene binding-site detector 118 is to detect a BINDING SITE of Fig. 8, located in an untranslated region of the RNA of a known gene, the nucleotide sequence of which BINDING SITE is at least partially complementary to that of a GAM RNA of Fig. 8, thereby determining that the above-mentioned known gene is a target gene of GAM of Fig. 8.

[0169] The target-gene binding-site detector 118 of Fig. 9 therefore receives as input a plurality of dicer-cut sequences from hairpins 140 of Fig. 13A which were calculated by the previous step, and a plurality of potential target gene sequences 142 which derive sequence DNA data 104 of Fig. 9, and after a phase of target-gene binding-site detector training & validation 128 of Fig. 10, is operative to detect target-genes having binding site/s 144 the nucleotide sequence of which is at least partially complementary to that of each of the plurality of dicer-cut sequences from hairpins 140. Preferred operation of the target-gene binding-site detector is further described hereinbelow with reference to Fig. 14B.

[0170] Reference is now made to Fig. 14B which is a simplified flowchart illustrating a preferred operation of the target-gene binding-site detector 118 of Fig. 9. In a preferred embodiment of the present invention, the target-gene binding-site detector 118 first performs a BLAST comparison of the nucleotide sequence of each of the plurality of dicer-cut sequences from hairpins 140, to the potential target gene sequences 142, in order to find crude potential matches. Blast results are then filtered to results which are similar to those of known binding sites (e.g. binding sites of miRNA genes Lin-4 and Let-7 to target genes Lin-14, Lin-41, Lin 28 etc.). Next the binding site is expanded, checking if nucleotide sequenced immediately adjacent to the binding site found by BLAST, may improve the match. Suitable binding sites, then are computed for free-energy and spatial structure. The results are analyzed, selecting only those binding sites, which have free-energy and spatial structure similar to that of known binding sites.

[0171] Reference is now made to Fig. 15 which is a simplified flowchart illustrating a preferred operation of the function & utility analyzer 120 described hereinabove with reference to Fig. 9. The goal of the function & utility analyzer

120 is to determine if a potential target gene is in fact a valid clinically useful target gene. Since a potential novel GAM gene binding a binding site in the UTR of a target gene is understood to inhibit expression of that target gene, and if that target gene is shown to have a valid clinical utility, then in such a case it follows that the potential novel gene itself also has a valid useful function which is the opposite of that of the target gene.

[0172] The function & utility analyzer 120 preferably receives as input a plurality of potential novel target genes having binding-site/s 144, generated by the target-gene binding-site detector 118, both of Fig. 14A. Each potential gene, is evaluated as follows:

[0173] First the system first checks to see if the function of the potential target gene is scientifically well established. Preferably, this can be achieved bioinformatically by searching various published data sources presenting information on known function of proteins. Many such data sources exist and are published as is well known in the art.

[0174] Next, for those target genes the function of which is scientifically known and is well documented, the system then checks if scientific research data exists which links them to known diseases. For example, a preferred embodiment of the present invention utilizes the OMIM(TM) database published by NCBI, which summarizes research publications relating to genes which have been shown to be associated with diseases.

[0175] Finally, the specific possible utility of the target gene is evaluated. While this process too may be facilitated by bioinformatic means, it might require human evaluation of published scientific research regarding the target gene, in order to determine the utility of the target gene to the diagnosis and or treatment of specific disease. Only potential novel genes, the target-genes of which have passed all three examinations, are accepted as novel genes.

[0176] Reference is now made to Fig. 16, which is a simplified diagram describing a novel bioinformatically detected group of regulatory genes, referred to here as Genomic Record (GR) genes, that encode an "operon-like" cluster of novel miRNA-like genes, each modulating expression of a plurality of target genes, the function and utility of which target genes is known.

[0177] GR GENE (Genomic Record Gene) is gene of a novel, bioinformatically detected group of regulatory, non protein coding, RNA genes. The method by which GR is detected is described hereinabove with reference to Figs. 8-18.

[0178] GR GENE encodes an RNA molecule, typically several hundred nucleotides long, designated GR PRECURSOR RNA.

[0179] GR PRECURSOR RNA folds spatially, as illustrated by GR FOLDED PRECURSOR RNA, into a plurality of what is known in the art as "hairpin" structures. The nucleotide sequence of GR PRECURSOR RNA comprises a plurality of segments, the first half of each such segment having a nucleotide sequence which is at least a partial inversed-reversed sequence of the second half thereof, thereby causing formation of a plurality of "hairpin" structures, as is well known in the art.

[0180] GR FOLDED PRECURSOR RNA is naturally processed by cellular enzymatic activity, into three separate hairpin shaped RNA segments, each corresponding to GAM PRECURSOR RNA of Fig. 8, designated GAM1 FOLDED PRECURSOR, GAM2 FOLDED PRECURSOR and GAM3 FOLDED PRECURSOR respectively.

[0181] The above mentioned GAM precursors, are diced by DICER COMPLEX of Fig. 8, yielding short RNA segments of about 22 nucleotides in length, each corresponding to GAM RNA of Fig. 8, designated GAM1 RNA, GAM2 RNA and GAM3 RNA respectively.

[0182] GAM1 RNA, GAM2 RNA and GAM3 RNA each bind complementarily to binding sites located in untranslated regions of respective target genes, designated GAM1-TARGET RNA, GAM2-TARGET RNA and GAM3-TARGET RNA respectively. This binding inhibits translation of the respective target proteins designated GAM1-TARGET PROTEIN, GAM2-TARGET PROTEIN and GAM3-TARGET PROTEIN respectively.

[0183] The structure of GAM genes comprised in a GR GENE, and their mode of modulation of expression of their respective target genes is described hereinabove with reference to Fig. 8. The bioinformatic approach to detection of GAM genes comprised in a GR GENE is described hereinabove with reference to Figs. 8 through 18. The present invention discloses 7,399 novel genes of the GR group of genes, which have been detected bioinformatically, as described hereinbelow. Laboratory confirmation of 3 genes of the GR group of genes is described hereinbelow with reference to Figs. 21 through 23.

[0184] Detailed descriptions of each of a plurality of GR GENEs and a detailing of GAM GENEs comprised in each of a plurality of GR GENEs of Fig. 16 is provided in Table 8, hereby incorporated by reference.

[0185] Nucleotide sequences of each said GAM GENEs and their respective genomic source and chromosomal location are further described hereinbelow with reference to Table 2 through Table 4, hereby incorporated by reference. GAM TARGET GENEs of each of said GAM GENEs are elaborated hereinbelow with reference to Table 5, hereby incorporated by reference. The functions of each of said GAM TARGET GENEs and their association with various diseases, and accordingly the utilities of said each of GAM GENEs, and hence the functions and utilities of each of said GR GENEs of Fig. 16 is elaborated hereinbelow with reference to Table 6, hereby incorporated by reference. Studies establishing known functions of each of said GAM TARGET GENEs, and correlation of each of said GAM TARGET GENEs to known diseases are listed in Table 7, and are hereby incorporated by reference.

[0186] In summary, the current invention discloses a very large number of novel GR genes, each of which encodes a plurality of GAM genes, which in turn may modulate expression of a plurality of target proteins. It is appreciated therefore that the function of GR genes is in fact similar to that of the Genomic Records concept of the present invention addressing the differentiation enigma, described hereinabove with reference to Fig. 7.

[0187] Reference is now made to Fig. 17 which is a simplified diagram illustrating a mode by which genes of a novel group of operon-like genes, described hereinabove with reference to Fig. 16 of the present invention, modulate expression of other such genes, in a cascading manner.

[0188] GR1 GENE and GR2 GENE are two genes of the novel group of operon-like genes designated GR of Fig. 16. As is typical of genes of the GR group of genes, GR1 and GR2 each encode a long RNA precursor, which in turn folds into a folded RNA precursor comprising multiple hairpin shapes, and is cut into respective separate hairpin shaped RNA segments, each of which RNA segments being diced to yield a gene of a group of genes designated GAM RNA of Fig. 8. In this manner GR1 yields GAM1, GAM2 and GAM3, and GR2 yields GAM4, GAM5 and GAM6.

[0189] As Fig. 17 shows, GAM 3 which derives from GR1, binds a binding site located adjacent to GR2 GENE, thus modulating expression of GR2, thereby invoking expression of GAM4, GAM5 and GAM6 which derive from GR2.

[0190] It is appreciated that the mode of modulation of expression presented by Fig. 17 enables an unlimited "cascading effect" in which a GR gene comprises multiple GAM genes, each of which may modulate expression of other GR genes, each such GR gene comprising additional GAM genes, etc., whereby eventually certain GAM genes modulate expression of target proteins. This mechanism is in accord with the conceptual model of the present invention addressing the differentiation enigma, described hereinabove with specific reference to Figs. 6 and 7.

[0191] Reference is now made to Fig. 18 which is a block diagram illustrating an overview of a methodology for finding novel genes and operon-like genes of the present invention, and their respective functions.

[0192] According to a preferred embodiment of the present invention, the methodology to finding novel genes of the present invention and their function comprises of the following major steps:

[0193] First, genes of the novel group of genes of the present invention, referred to here as GAM genes, are located and their function elicited by detecting target proteins they bind and the function of those target proteins, as described hereinabove with reference to Figs. 8 through 15.

[0194] Next, genes of a novel group of operon-like genes of the present invention, referred to here as GR genes, are located, by locating clusters of proximally located GAM genes, based on the previous step.

[0195] Consequently, the hierarchy of GR and GAM genes is elicited: binding sites for non-protein-binding GAM genes comprised in each GR gene found, are sought adjacent to other GR genes. When found, such a binding site indicates that the connection between the GAM and the GR the expression of which it modulates, and thus the hierarchy of the GR genes and the GAM genes they comprise.

[0196] Lastly, the function of GR genes and GAM genes which are "high" in the hierarchy, i.e. GAM genes which modulate expression of other GR genes rather than directly modulating expression of target proteins, may be deduced. A preferred approach is as follows: The function of protein-modulating GAM genes is deducible from the proteins which they modulate, provided that the function of these target proteins are known. The function of "higher" GAM genes may be deduced by comparing the function of protein-modulating GAM genes, with the hierarchical relationships by which the "higher" GAM genes are connected to the protein-modulating GAM genes. For example, given a group of several protein-modulating GAM genes, which collectively cause a protein expression pattern typical of a certain cell-type, then a "higher" GAM gene is sought which modulates expression of GR genes which perhaps modulate expression of other genes which eventually modulate expression of the given group of protein-modulating GAM genes. The "higher" GAM gene found in this manner, is taken to be responsible for differentiation of that cell-type, as per the conceptual model of the invention described hereinabove with reference to Fig. 6.

[0197] Reference is now made to Fig. 19 which is a block diagram illustrating different utilities of genes of the novel group of genes of the present invention referred to here as GAM genes and GR genes.

[0198] The present invention discloses a first plurality of novel genes referred to here as GAM genes, and a second plurality of operon-like genes referred to here as GR genes, each of the GR genes encoding a plurality of GAM genes. The present invention further discloses a very large number of known target-genes, which are bound by, and the expression of which is modulated by each of the novel genes of the present invention. Published scientific data referenced by the present invention provides specific, substantial, and credible evidence that the abovementioned target genes modulated by novel genes of the present invention, are associated with various diseases.

Specific novel genes of the present invention, target genes thereof and diseases associated therewith, are described hereinbelow with reference to Figs. 23 through 30,022. It is therefore appreciated that a function of GAM genes and GR genes of the present invention is modulation of expression of target genes related to known diseases, and that therefore utilities of novel genes of the present invention include diagnosis and treatment of the abovementioned diseases. Fig. 19 describes various types of diagnostic and therapeutic utilities of novel genes of the present invention.

[0199] A utility of novel genes of the present invention is detection of GAM genes and of GR genes. It is appreciated that since GAM genes and GR genes modulate expression of disease related target genes, that detection of expression of GAM genes in clinical scenarios associated with said diseases is a specific, substantial and credible utility. Diagnosis of novel genes of the present invention may preferably be implemented by RNA expression detection techniques, including but not limited to biochips, as is well known in the art. Diagnosis of expression of genes of the present invention may be useful for research purposes, in order to further understand the connection between the novel genes of the present invention and the abovementioned related diseases, for disease diagnosis and prevention purposes, and for monitoring disease progress.

[0200] Another utility of novel genes of the present invention is anti-GAM gene therapy, a mode of therapy which allows up regulation of a disease related target-gene of a novel GAM gene of the present invention, by lowering levels of the novel GAM gene which naturally inhibits expression of that target gene. This mode of therapy is particularly useful with respect to target genes which have been shown to be under-expressed in association with a specific disease. Anti-GAM gene therapy is further discussed hereinbelow with reference to Figs. 20A and 20B.

[0201] A further utility of novel genes of the present invention is GAM replacement therapy, a mode of therapy which achieves down regulation of a disease related target-gene of a novel GAM gene of the present invention, by raising levels of the GAM gene which naturally inhibits expression of that target gene. This mode of therapy is particularly useful with respect to target genes which have been shown to be over-expressed in association with a specific disease. GAM replacement therapy involves introduction of supplementary GAM gene products into a cell, or stimulation of a cell to produce excess GAM gene products. GAM replacement therapy may preferably be achieved by transfecting cells with an artificial DNA molecule encoding a GAM gene, which causes the cells to produce the GAM gene product, as is well known in the art.

[0202] Yet a further utility of novel genes of the present invention is modified GAM therapy. Disease conditions are likely to exist, in which a mutation in a binding site of a GAM gene prevents natural GAM gene to effectively bind inhibit a disease related target-gene, causing up regulation of that target gene, and thereby contributing to the disease pathology. In such conditions, a modified GAM gene is designed which effectively binds the mutated GAM binding site, i.e. is an effective anti-sense of the mutated GAM binding site, and is introduced in disease effected cells. Modified GAM therapy is preferably achieved by transfecting cells with an artificial DNA molecule encoding the modified GAM gene, which causes the cells to produce the modified GAM gene product, as is well known in the art.

[0203] An additional utility of novel genes of the present invention is induced cellular differentiation therapy. As aspect of the present invention is finding genes which determine cellular differentiation, as described hereinabove with reference to Fig. 18. Induced cellular differentiation therapy comprises transfection of cell with such GAM genes thereby determining their differentiation as desired. It is appreciated that this approach may be widely applicable, inter alia as a means for auto transplantation harvesting cells of one cell-type from a patient, modifying their differentiation as desired, and then transplanting them back into the patient. It is further appreciated that this approach may also be utilized to modify cell differentiation in vivo, by transfecting cells in a genetically diseased tissue with a cell-differentiation determining GAM gene, thus stimulating these cells to differentiate appropriately.

[0204] Reference is now made to Figs. 20A and 20B, simplified diagrams which when taken together illustrate anti-GAM gene therapy mentioned hereinabove with reference to Fig. 19. A utility of novel genes of the present invention is anti-GAM gene therapy, a mode of therapy which allows up regulation of a disease related target-gene of a novel GAM gene of the present invention, by lowering levels of the novel GAM gene which naturally inhibits expression of that target gene. Fig. 20A shows a normal GAM gene, inhibiting translation of a target gene of GAM gene, by binding to a BINDING SITE found in an untranslated region of GAM TARGET RNA, as described hereinabove with reference to Fig. 8.

[0205] Fig. 20B shows an example of anti-GAM gene therapy. ANTI-GAM RNA is short artificial RNA molecule the sequence of which is an anti-sense of GAM RNA. Anti-GAM treatment comprises transfecting diseased cells with ANTI-GAM RNA, or with a DNA encoding thereof. The ANTI-GAM RNA binds the natural GAM RNA, thereby preventing binding of natural GAM RNA to its BINDING SITE. This prevents natural translation inhibition of TARGET RNA by GAM RNA, thereby up regulating expression of GAM TARGET PROTEIN.

[0206] It is appreciated that anti-GAM gene therapy is particularly useful with respect to target genes which have been shown to be under-expressed in association with a specific disease.

[0207] Reference is now made to Fig. 21A which is an annotated sequence of an EST comprising a novel gene detected by the gene detection system of the present invention. Fig. 21A shows the nucleotide sequence of a known human non-protein coding EST (Expressed Sequence Tag), identified as EST72223. It is appreciated that the sequence of this EST comprises sequences of one known miRNA gene, identified as MIR98, and of one novel GAM gene, referred to here as GAM25, detected by the bioinformatic gene detection system of the present invention, described hereinabove with reference to Fig. 9.

[0208] Reference is now made to Figs. 21B and 21C that are pictures of laboratory results, which when taken together demonstrate laboratory confirmation of expression of the bioinformatically detected novel gene of Fig. 21A. Reference is now made to Fig. 21B which is a Northern blot analysis of MIR-98 and EST72223 transcripts. MIR-98 and EST72223 were reacted with MIR-98 and GAM25 probes as indicated in the figure. It is appreciated that the probes of both MIR-98 and GAM25 reacted with EST72223, indicating that EST72223 contains the sequences of MIR-98 and of GAM25. It is further appreciated that the probe of GAM25 does not cross-react with MIR-98.

[0209] Reference is now made to Fig. 21C. A Northern blot analysis of EST72223 and MIR-98 transfections were performed, subsequently marking RNA by the MIR-98 and GAM25 probes . Left, Northern reacted with MIR-98, Right, Northern reacted with GAM25. The molecular Sizes of EST72223, MIR-98 and GAM25 are indicated by arrows.

Hela are control cells that have not been introduced to exogenous RNA. EST and MIR-98 Transfections are RNA obtained from Hela transfected with EST72223 and MIR-98, respectively. MIR-98 and EST are the transcripts used for the transfection experiment. The results indicate that EST72223, when transfected into Hela cells, is cut yielding known miRNA gene MIR-98 and novel miRNA gene GAM25.

[0210] Reference is now made to Fig. 21D, which is a Northern blot of a lysate experiment with MIR-98 and GAM25. Northern blot analysis of hairpins in EST72223 . Left, Northern reacted with predicted Mir-98 hairpin probe, Right, Northern reacted with predicted GAM25 hairpin probe. The molecular size of EST Is indicated by arrow. The molecular sizes of Mir-98 and GAM25 are 80nt and 100nt, respectively as indicated by arrows. The 22nt molecular marker is indicated by arrow. 1-Hela lysate; 2-EST incubated 4h with Hela lysate; 3-EST without lysate; 4-Mir transcript incubated 4h with Hela lysate; 5-Mir transcript incubated overnight with Hela lysate; 6- Mir transcript without lysate; 7-RNA extracted from Hela cells following transfection with Mir transcript.

[0211] Technical methods used in experiments, the results of which are depicted in Figs. 21B, 21C and 21D are as follows:

[0212] *Transcript preparations:* Transcripts were prepared of EST72223(TIGR) and of MIR98 and predicted GAM4 within it, and of EST7929020(IMAGE) and of predicted GAM3 within it.. Transcripts were prepared by m^{7}G(5')ppp(5')G-capping reaction by using mMessage mMachine kit (Ambion) according to the manufacture"s protocol. Briefly, PCR products amplified with specific primers contain T7 promoter at the 5" end and T3 promoter at the 3"end were prepared from each DNA. The purified PCR products were transcribed with T7 polymerase. Transcript products were 725nt (EST72223), 102nt (MIR98), 125nt (GAM4) and 70nt (GAM3) long.EST72223 was PCR amplified with T7-EST 72223 forward primer:5"-TAATACGACTCACTATAGGCCCTTATTAGAGGATTCTGCT-3" and T3-EST72223 reverse primer: 5"-AATTAACCCTCACTAAAGGTTTTTTTTTCCTGAGACAGAGT-3".MIR98 was PCR amplified using EST72223 as a template with T7MIR98 forward primer: 5-"TAATACGACTCACTATAGGGTGAGGTAGTAAGTTGTATTGTT-3"and T3MIR98 reverse primer:

5"-AATTAACCCTCACTAAAGGAAAGTAGTAAGTTGTATAG TT-3".GAM4 was PCR amplified using EST72223 as a template with GAM4 forward primer: 5"-GAGGCAGGAGAATTGCTTGA- 3" and T3-EST72223 reverse primer:5"-AATTAACCCTCACTAAAGGCCTGAGACAGAGTCT TGCTC-3".GAM3 was PCR amplified using EST7929020 as a template with T7-GAM3 forward primer: 5"-TAATACGACTCACTATAGGGTCAGAGTGAACAGGCAACC -3" and T3-GAM3 reverse primer:5"-AATTAACCCTCACTAAAGGGTCAGATGAGTAGGT TGCGAA -3".

[0213] *Transfection procedure:* Capped RNA transcripts were incubated at 30°C in supplemented Hela S100 obtained from 4C Biotech, Seneffe, Belgium. The Hela S100 was supplemented by dialysis to a final concentration of 20mM Hepes, 100mM KCl, 2.5mM $MgCl_2$, 0.5mM DTT, 20% glycerol and protease inhibitor cocktail tablets (Complete mini Roche Molecular Biochemicals) and kept at 4°C for one month. After addition of all components, final concentrations were 100mM capped target RNA, 2mM ATP, 0.2mM GTP, 500U/ml RNasin, 30µg/ml creatine kinase, 25mM creatine phosphate, 2.5mM DTT and 50% S100 extract. Cleavage reaction was stopped at different time points (0, 0.5, 1, 4, 24h) by the addition of 8 volumes of proteinase K buffer (200Mm Tris-Hcl, pH 7.5, 25m M EDTA, 300mM NaCl, and 2% SDS) and incubated at 65°C for 15min. Proteinase K, dissolved in 50mM Tris-HCl, pH 8, 5m M $CaCl_2$, and 50% glycerol, was added to a final concentration of 0.6 mg/ml. Samples were subjected to phenol/chloroform extraction. Pellets were dissolved in water and kept frozen. Samples were analyzed after addition of two volumes TBE-Urea buffer (1xTBE, 7M urea) on a TBE-Urea PAGE gel.

[0214] *Target RNA cleavage assay*: Digoxigenin (DIG) labeled antisense transcripts was prepared from purified PCR product of MIR98 by using a DIG RNA labeling kit with T3 polymerase (Roche Molecular Biochemicals) according to the manufacturer"s protocol. PCR primers are detailed above. Labeled transcript was 102nt long.Digoxigenin (DIG) labeled PCR was prepared for GAM4 by using a DIG PCR labeling kit (Roche Molecular Biochemicals) according to the manufacture"s protocol. PCR primers are detailed above. Labeled PCR was 145bp long.3"-DIG-tailed oligo ssDNA antisense probes, containing DIG-dUTP and dATP at an average tail length of 50 nucleotides were prepared with the DIG Oligonucleotide Labeling Kit (Roche Molecular Biochemicals) from 100pmole oligonucleotides. Labeled predicted "correct"GAM3 oligonucleotide is 5"-GAGTAGGTTGCGAAAATTTTCTCC-3" and labeled "incorrect"GAM3 oligonucleotide is 5"-CCCATTTTGTAGGTTGCCTGTTCA-3", and predicted "correct"GAM4 oligonucleotide is 5"-CTTCCTGGGTTCAAGCAATT-3", while labeled "incorrect"GAM4 oligonucleotide is 5"-. CTGAGACA-GAGTCTTGCTCTG-3". Labeled oligo probes were 24nt long.

[0215] *Northern analysis:* RNA samples were boiled for 3 min before loading on a segmented, top 6%, bottom 13% polyacrylamide gel, containing 7M urea and 1xTBE. Gels were run in 1xTBE at a constant voltage of 250V and then transferred onto a positively charged nylon membrane (Biodyne PLUS 0.45µm, Pall) and UV cross-linked. Hybridization was performed overnight with DIG-labeled probes at 42^0C in DIG Easy-Hyb buffer (Roche). Membranes were washed twice with 2xSSC and 0.1% SDS for 10 min. at 42^0C and then washed twice with 0.5xSSC and 0.1% SDS for 5 min at 42^0C. The membrane was then developed by using a DIG luminescent detection kit (Roche) using anti-DIG and CSPD reaction, according to the manufacturer"s protocol.

[0216] It is appreciated that the data presented in Figs. 21A, 21B, 21C and 21D, when taken together validate the function of the bioinformatic gene detection engine 100 of Fig. 9. Fig. 21A shows a novel GAM gene bioinformatically detected by the bioinformatic gene detection engine 100, and Figs. 21B, 21C and 21D show laboratory confirmation of the expression of this novel gene. This is in accord with the engine training and validation methodology described hereinabove with reference to Fig. 10.

[0217] Reference is now made to Fig. 22A which is an annotated sequence of an EST comprising a novel gene detected by the gene detection system of the present invention. Fig. 22A shows the nucleotide sequence of a known human non-protein coding EST (Expressed Sequence Tag), identified as EST 7929020. It is appreciated that the sequence of this EST comprises sequences of two novel GAM genes, referred to here as GAM24 and GAM26, detected by the bioinformatic gene detection system of the present invention, described hereinabove with reference to Fig. 9.

[0218] Reference is now made to Fig. 22B which presents pictures of laboratory results, that demonstrate laboratory confirmation of expression of the bioinformatically detected novel gene of Fig. 22A. Northern blot analysis of hairpins in EST7929020. Left, Northern reacted with predicted GAM26 hairpin probe, Right, Northern reacted with predicted GAM24 hairpin probe. The molecular size of EST is indicated by arrow. The molecular sizes of GAM24 and GAM26 are 60nt, as indicated by arrow. The 22nt molecular marker is indicated by arrow. 1-Hela lysate; 2- EST incubated 4h with Hela lysate ; 3- EST incubated overnight with Hela lysate; 4-EST without lysate; 5-GAM transcript; 6- GAM 22nt marker;7-GAM PCR probe; 8-RNA from control Hela cells; 9-RNA extracted from Hela cells following transfection with EST.

[0219] Reference is now made to Fig. 22C which is a picture of a Northern blot confirming Endogenous expression of bioinformatically detected gene GAM25 of Fig. 22A from in Hela cells. Northern was reacted with a predicted GAM26 hairpin probe. The molecular size of EST7929020 is indicated. The molecular sizes of GAM26 is 58nt, as indicated. A 19nt DNA oligo molecular marker is indicated. Endogenous expression of GAM26 in Hela total RNA fraction and in S-100 fraction is indicated by arrows. 1-GAM26 transcript; 2- GAM25 DNA oligo marker; 3-RNA from control Hela cells; 4-RNA extracted from Hela cells following transfection with EST; 5- RNA extracted from S-100 Hela lysate.

[0220] Reference is now made to Fig. 23A which is an annotated sequence of an EST comprising a novel gene detected by the gene detection system of the present invention. Fig. 23A shows the nucleotide sequence of a known human non-protein coding EST (Expressed Sequence Tag), identified as EST 1388749. It is appreciated that the sequence of this EST comprises sequence of a novel GAM gene, referred to here as GAM27, detected by the bioinformatic gene detection system of the present invention, described hereinabove with reference to Fig. 9.

[0221] Reference is now made to Fig. 23B which is a picture of Northern blot analysis, confirming expression of novel bioinformatically detected gene GAM26, and natural processing thereof from EST1388749. Northern reacted with predicted GAM27 hairpin probe. The molecular size of EST is indicated by arrow. The molecular sizes of GAM27 is 130nt, as indicated by arrow. The 22nt molecular marker is indicated by arrow. 1-Hela lysate; 2-EST incubated 4h with Hela lysate; 3- EST incubated overnight with Hela lysate; 4-EST without lysate; 5-GAM transcript; 6- GAM

22nt marker; 7-GAM PCR probe.

[0222] It is appreciated by persons skilled in the art that the present invention is not limited by what has been particularly shown and described hereinabove. Rather the scope of the present invention includes both combinations and subcombinations of the various features described hereinabove as well as variations and modifications which would occur to persons skilled in the art upon reading the specifications and which are not in the prior art.

DETAILED DESCRIPTION OF LARGE TABLES

[0223] Table 1 comprises detailed textual description according to the description of Fig.8 of each of a plurality of novel GAM gene of the present invention, and contains the following fields: GENE NAME:Rosetta Genomics Ltd. gene nomenclature (see below);PRECUR SEQ-ID:GAM precursor Seq-ID, as in the Sequence Listing;PRECURSOR SEQUENCE: Sequence (5` to 3`) of the GAM precursor gene; GENE DESCRIPTION: Detailed description of GAM gene with reference to Fig.8; and

[0224] Table 2 comprises data relating to the source and location of novel GAM genes of the present invention, and contains the following fields: GENE NAME: Rosetta Genomics Ltd. gene nomenclature (see below); PRECUR SEQ-ID: GAM precursor Seq-ID, as in the Sequence Listing; ORGANISM: Abbreviated (hsa = Homo sapiens); CHR: Chromosome encoding the GAM gene; STRAND: Orientation on the chromosome, '+' represents the plus strand, '-' represents the minus strand; CHR-START OFFSET: Start offset of GAM precursor sequence on chromosome; CHR-END OFFSET: End offset of GAM precursor sequence on chromosome; SOURCE REF-ID: Accession number of source sequence; and

[0225] Table 3 comprises data relating to GAM precursors of novel GAM genes of the present invention, and contains the following fields:GENE NAME: Rosetta Genomics Ltd. gene nomenclature (see below); PRECUR SEQ-ID: GAM precursor Seq-ID, as in the Sequence Listing; PRECURSOR SEQUENCE: GAM precursor nucleotide sequence (5\` to 3\`); GAM FOLDED-PRECURSOR: Schematic representation of the GAM folded precursor, beginning 5\` end (beginning of upper row) to 3\` end (beginning of lower row), where the hairpin loop is positioned at the right part of the draw; and

[0226] Table 4 comprises data relating to GAM genes of the present invention, and contains the following fields: GENE NAME: Rosetta Genomics Ltd. gene nomenclature (see below); GAM SEQ-ID: GAM Seq-ID, as in the Sequence Listing; GENE SEQUENCE: Sequence (5` to 3`) of the mature, `diced` GAM gene; PRECUR SEQ-ID : GAM precursor Seq-ID, as in the Sequence Listing; SOURCE REF-ID: Accession number of the source sequence; GAM POS: Dicer cut location (see below);and

[0227] Table 5 comprises data relating to target-genes and binding sites of GAM genes of the present invention, and contains the following fields: GENE NAME: Rosetta Genomics Ltd. gene nomenclature (see below); GAM SEQ-ID: GAM Seq-ID, as in the Sequence Listing; TARGET: GAM target gene name; #BS: Number of unique binding sites of GAM onto Target utr side ; TARGET SEQ-ID: Target binding site Seq-ID, as in the Sequence Listing; UTR: Untranslated region of binding site/s (3" or 5"); TARGET BINDING SITE-SEQ: Nucleotide sequence (5` to 3`) of the target binding site; BINDING-SITE DRAW: Schematic representation of the binding site, upper row present 5` to 3` sequence of the GAM, lower row present 3` to 5` sequence of the target; GAM POS:Dicer cut location (see below);and

[0228] Table 6 comprises data relating to functions and utilities of novel GAM genes of the present invention, and contains the following fields: GENE NAME: Rosetta Genomics Ltd.

gene nomenclature (see below);TARGET: GAM target gene name; GENE SEQUENCE: Sequence (5` to 3`) of the mature, `diced` GAM gene; GENE FUNCTION: Description of the GAM functions and utilities; GAM POS: Dicer cut location (see below);TAR DIS: Target Disease Relation Group (see below); and

[0229] Table 7 comprises data of gene function references - Bibliography and contains the following fields: GENE NAME: Rosetta Genomics Ltd. gene nomenclature (see below);GAM SEQ-ID: GAM Seq-ID, as in the Sequence Listing; TARGET: GAM target gene name; REFERENCES: list of references relating to the target gene; GAM POS: Dicer cut location (see below); and

[0230] Table 8 comprises data relating to novel GR genes of the present invention, and contains the following fields: GENE NAME: Rosetta Genomics Ltd. GR gene nomenclature; SOURCE REF-ID: Accession number of the source sequence; GR DESCRIPTION: Detailed description of a GR gene cluster, with reference to Fig.16; SRC: Source-accession number of GR sequence ; and

[0231] The following conventions and abbreviations are used in the tables: The nucleotide 'U' is represented as 'T' in the tables.

[0232] GENE NAME is a RosettaGenomics Ltd. gene nomenclature. All GAMs are designated by GAMx where x is a unique ID number.

[0233] GAM POS is a position of the GAM RNA on the GAM PRECURSOR RNA sequence. This position is the Dicer cut location, A indicates a probable Dicer cut location, B indicates an alternative Dicer cut location.

[0234] TAR DIS (Target Disease Relation Group) 'A' indicates if the target gene is known to have a specific causative relation to a specific known disease, based on the OMIM database. It is appreciated that this is a partial classification emphasizing genes which are associated with "single gene" diseases etc. All genes of the present invention ARE associated with various diseases, although not all are in "A" status.